Week 7: Linear Regression model and Goodness criteria

Ziwen Ye

Stevens Institute of Technology

zye2@stevens.edu

October 13, 2018

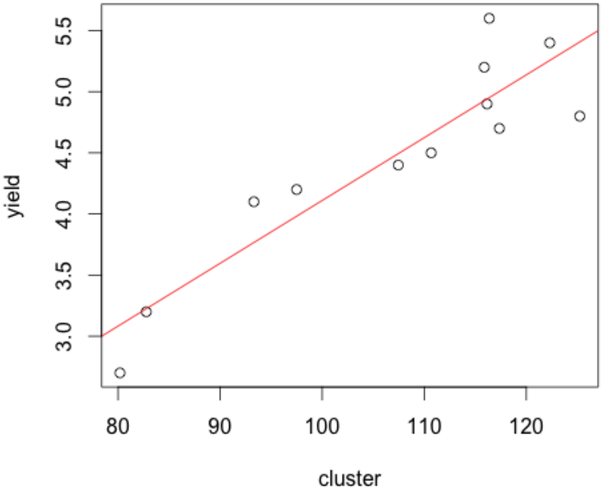
Outline

1. Linear Regression

2. Multiple Regression

3. Stepwise regression and goodness criteria

Simple Linear Regression



Simple Linear Regression

Objective

Describe the relationship between two variables, say X and Y as a straight line, that is, Y is modeled as a linear function of X.

Variables

X: explanatory variable

Y: response variable

After data collection, we have pairs of observations:

$$(x_1, y_1), ..., (x_n, y_n)$$

Simple Linear Regression

Model

Then we fit our data set into this simple linear model.

$$Y = \alpha + \beta X + \epsilon, \text{ where } i = 1, 2, ..., n$$

- Residuals: $\epsilon_i \sim N(0, \sigma^2)$, independent
- Estimate y_j by the value of x_j, $\hat{y}_j = \alpha + \beta x_j = E(y_j|x_j)$

Parameters

We need to figure out a way to estimate these parameters:

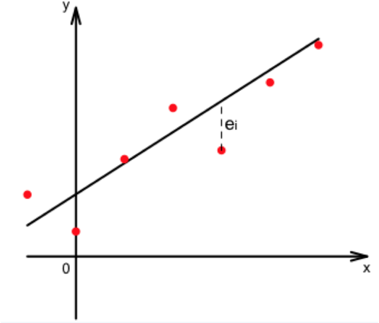
- α (Intercept): point in which the line intercepts the y-axis
- β (Slope): increase in Y per unit change in X

Least Squares

We want to find the equation of the line that best fits the data. It means finding α and β such that the fitted values of y_j , given by

$$\hat{y}_j = \alpha + \beta x_j$$

are as close as possible to the observed values y_i, for all $i = 1, 2, ..., n$.



residuals given by:
$\epsilon_i = y_i - \hat{y}_i$

Least Squares

Estimation of Parameters

A usual way of calculating α and β is based on the minimization of the sum of the squared residuals, or residual sum of squares (RSS):

$$\begin{aligned} RSS &= \Sigma_{i=1}^{n}\epsilon_i^2 \\ &= \Sigma_{i=1}^{n}(y_i - \hat{y}_i)^2 \\ &= \Sigma_{i=1}^{n}(y_i - \alpha - \beta x_i)^2 \end{aligned}$$

To solve this equation, we need to use gradient descent method. We will discuss more about it in future.

Simple Linear Regression

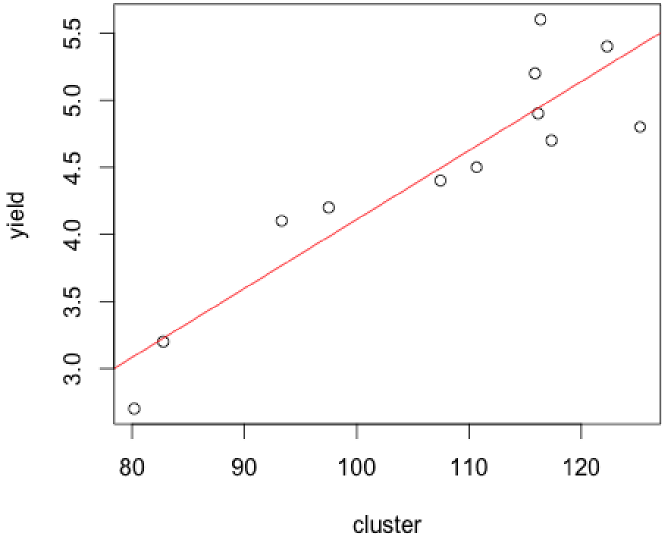
Example

```
> yield <- data$yield
> cluster <- data$cluster.count
> plot(yield ~ cluster)
> lm(yield ~ cluster)

Call:
lm(formula = yield ~ cluster)

Coefficients:
(Intercept)      cluster
   -1.02790      0.05138

> lm.r <- lm(yield ~ cluster)
> abline(lm.r, col=’red’)
```


yield
3.0
3.5
4.0
4.5
5.0
5.5
80
90
100
110
120
cluster

Example

```
> summary(lm.r)
Call:
lm(formula = yield ~ cluster)

Residuals:
     Min       1Q   Median       3Q      Max
-0.60700 -0.19471 -0.03241  0.23220  0.64874

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.02790    0.78355  -1.312    0.219
cluster      0.05138    0.00725   7.087 3.35e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
Residual standard error: 0.3641 on 10 degrees of freedom
Multiple R-squared:  0.834,Adjusted R-squared:  0.8174
F-statistic: 50.23 on 1 and 10 DF,  p-value: 3.347e-05
```

Reading summary for a linear model

Estimation

The estimation gives you the value of estimated parameter. Based on the summary report, $\alpha = -1.027$ and $\beta = 0.051$

P-values

P-value is the index, which helps you to understand whether the factor is significant or not in your model. In most cases, only when the P-value is less than 0.05 (or ***), we think this factor is significant.

Adjusted R square values

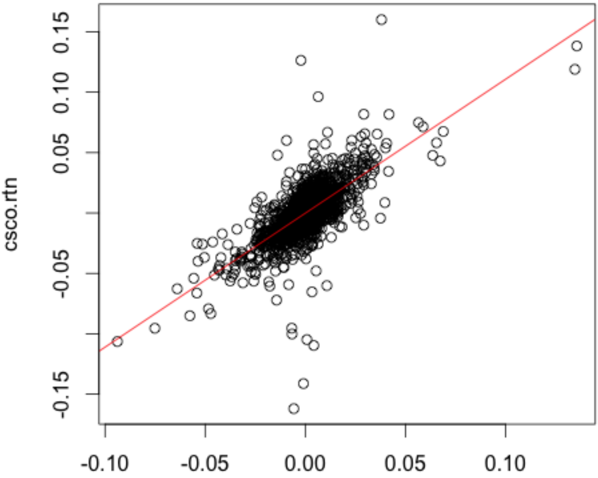
In the summary report, R-squared value is the most important index. The higher it is, the better your model is.

Example on Stock Returns

Example

```
> getSymbols("CSCO")
> getSymbols("DIA")
> csco <- data.frame(CSCO)
> dia <- data.frame(DIA)
> # get last price
> csco.price <- csco$CSCO.Adjusted
> dia.price <- dia$DIA.Adjusted
> # get returns
> csco.rtn <- diff(csco.price, lag = 1)/
            csco.price[-length(csco.price)]
> dia.rtn <- diff(dia.price, lag = 1)/
            dia.price[-length(dia.price)]
> plot(csco.rtn ˜ dia.rtn)
> lm1 <- lm(csco.rtn ˜ dia.rtn)
> abline(lm1, col = ’red’)
```

Example on Stock Returns



Different Types of Linear Models in R

Syntax	Model
$Y \sim X$	$Y = \alpha + \beta X$
$Y \sim -1 + X$	$Y = \beta X$
$Y \sim X_1 : X_2$	$Y = \alpha + \beta X_1 X_2$
$Y \sim X + I(X^2)$	$Y = \alpha + \beta_1 X + \beta_2 X^2$
$Y \sim X_1 + X_2$	$Y = \alpha + \beta_1 X_1 + \beta_2 X_2$
$Y \sim X_1 * X_2$	$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$

Multiple regression

Multiple linear regression model

In the multiple regression setting, the response variable y depends on more than one explanatory variables, which we denote by $x_1, x_2, ..., x_p$. The mean response depends on these explanatory variables according to a linear function

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + ...\beta_p x_p$$

Example

```
Call:
  lm(formula = voplus ~ vominus + oc + trap, data = bone)

Residuals:
  Min       1Q   Median      3Q      Max
-364.19 -158.57  -15.13  120.08  441.11

Coefficients:
  Estimate Std. Error t value Pr(>|t|)
(Intercept) -243.4877    94.2183  -2.584  0.01549 *
  vominus         0.9746     0.1211   8.048  1.2e-08 ***
  oc              8.2349     2.8397   2.900  0.00733 **
  trap            6.6071    10.3340   0.639  0.52797
  Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 207.8 on 27 degrees of freedom
Multiple R-squared:  0.8844,Adjusted R-squared:  0.8715
F-statistic: 68.84 on 3 and 27 DF,  p-value: 9.031e-13
```

Goodness criteria

Besides adjusted R square values, we also can use AIC or BIC.

AIC (Akaike information criterion)

- This is an estimate of the difference between the actual model and no model at all
- Small value indicate that the model is good

BIC (Bayesian information criterion)

- Similar to AIC, but penalize more in calculation.
- Small value indicate that the model is good

Remember, you can use any kind of goodness criteria when evaluating the model performance. However, you can only use one and stick with it throughout the question.

Stepwise Regression

Sometimes we may not sure about what variables should be used in the multiple linear regression. In such cases, we can use stepwise regression to delete variables which are not significant in the model (not useful at all or lurking variables)

- Forward selection (From 0 to n)
- Backward selection (From n to 0)
- Stepwise selection (Start with 0 and move to the next step based on the AIC value)